Ecotoxicology and Environmental Safety 82 (2012) 85–95

Contents lists available at SciVerse ScienceDirect

Ecotoxicology and Environmental Safety

journal homepage: www.elsevier.com/locate/ecoenv

Development of classification and regression based QSAR models to predict rodent carcinogenic potency using oral slope factor

Supratik Kar [a], Omar Deeb [b], Kunal Roy [a,*]

[a] Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
[b] Faculty of Pharmacy, Al-Quds University, P.O. Box 20002, Jerusalem, Palestine

ARTICLE INFO

Article history:
Received 9 January 2012
Received in revised form
6 February 2012
Accepted 21 May 2012
Available online 13 June 2012

Keywords:
OSF
Carcinogenicity
QSAR
OECD
LDA
PDD

ABSTRACT

Carcinogenicity is among the toxicological endpoints posing the highest concern for human health. Oral slope factors (OSFs) are used to estimate quantitatively the carcinogenic potency or the risk associated with exposure to the chemical by oral route. Regulatory agencies in food and drug administration and environmental protection are employing quantitative structure–activity relationship (QSAR) models to fill the data gaps related with properties of chemicals affecting the environment and human health. In this background, we have developed quantitative structure-carcinogenicity regression models for rodents based on the carcinogenic potential of 70 chemicals with wide diversity of molecular structures, spanning a large number of chemical classes and biological mechanisms. All the developed models have been assessed according to the Organization for Economic Cooperation and Development (OECD) principles for the validation of QSAR models. We have also attempted to develop a carcinogenicity classification model based on Linear Discriminant Analysis (LDA). Developed regression and LDA models are rigorously validated internally as well as externally. Our *in silico* studies make it possible to obtain a quantitative interpretation of the structural information of carcinogenicity along with identification of the discriminant functions between lower and higher carcinogenic compounds by LDA. Pharmacological distribution diagrams (PDDs) are used as a visualizing technique for the identification and selection of chemicals with lower carcinogenicity. Constructive, informative and comparable interpretations have been observed in both cases of classification and regression based modeling.

1. Introduction

To assess the potential risk of human carcinogens, long-term animal bioassays for carcinogenicity are regularly used to resolve whether chemical agents are proficient of inducing cancer in humans (CDER, 1997). According to the regulatory authorities of Europe, USA and Japan, carcinogenicity studies should be performed before the application for marketing approval of pharmaceuticals and chemicals (Müller et al., 1999). Guidelines for carcinogenicity testing of pharmaceuticals and chemicals specify that long-term carcinogenicity studies in rodents should be carried out to establish chemicals as a carcinogen (CDER, 1997). Rodent carcinogenicity studies have been used for many years to assess carcinogenic potential of chemicals with an ultimate goal of assessing human carcinogenic risk (Ward, 2010).

For evaluating the carcinogenic dose-response assessment and carcinogenic potency, one most commonly used measure is oral slope factor (OSF). The OSF is defined as an upper bound and it resembles a 95% confidence limit on the increased cancer risk from a lifetime exposure to a chemical or environmental contaminant (USEPA, 2009). The OSF is usually expressed in units of proportion affected per mg/kg/day. OSF is most commonly used by the Integrated Risk Information System (IRIS) of United States Environmental Protection Agency (US EPA) because OSF provides a linear extrapolation from the animal dose level to an environmental exposure level that is most relevant to human health (USEPA, 2009). OSF is generally reserved for use in the low-dose region of the dose-response relationship. On the contrary, tumor dose (TD_{50}) is not limited to the low-dose region and it is not true indicator of cancer from environmental exposures. Another significant variation is that TD_{50} does not provide a target organ-specific dose-response. On the other hand, OSF provides information on cancer at a target specific organ resulting from prolonged exposure to a chemical (USEPA, 2009).

The US EPA's IRIS provides information regarding health effects of chemicals to which the public may be exposed from

* Corresponding author. Fax: +91 33 2837 1078.
E-mail addresses: kunalroy_in@yahoo.com, kroy@pharma.jdvu.ac.in (K. Roy).
URL: http://sites.google.com/site/kunalroyindia/ (K. Roy).

0147-6513/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ecoenv.2012.05.013

releases to environment and through the use and disposal of chemicals. IRIS assessments provide a scientific foundation for decisions to protect public health across EPA's programs and regions under an array of environmental laws. Over the past two years, EPA has strengthened and smoothed the IRIS program, improving transparency and increasing the number of final assessments added to the IRIS database (IRIS Progress Report, 2011). The National Toxicology Program (NTP) plays a significant role in the identification and assessment of carcinogens in the US, while the International Agency for Research on Cancer (IARC) plays a vital role internationally (National Toxicology Program, 2005). Globally, the chemical industry and Regulatory Agencies such as the US EPA spend millions of dollars in testing and assessing the health risks associated with chemicals. But, it is difficult for U.S.EPA, other federal agencies and research organizations to make decisions regarding exposure guidelines for environmental contaminants when such experimental data are not available (USEPA, 2009). In such circumstances, *in silico* approaches, specifically quantitative structure-activity relationships (QSARs) have the ability to predict potential health hazards from chemical exposure through the use of developed correlation models. QSARs not only save time but also provide valuable resources which could be endowed more sensibly (Deeb, 2010; Kar and Roy, 2010). The European Chemical Bureau encourages the use of models in a regulatory framework while other agencies explicitly forbid the use of models for making regulatory decisions while allowing the use of correlations and models for screening assessments. Predictive models are used by Food and Drug Administration (FDA) to minimize false negatives and false positives saving incalculable costs for manufacturers (Benigni and Zito, 2004). On the other hand, increasing pressure from social and economic background to cut out the use of animal testing is another reason to develop alternative methods (European Commission, Directive 2006/121/EC), such as *in silico* QSAR models which also support 3Rs (replacement, refinement and reduction of animals in research) (Benigni and Giuliani, 2003) and Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) policies (Williams et al., 2009). However, the QSAR models should be validated according to Organization for Economic Cooperation and Development (OECD) principles for reliable prediction. These principles provide best possible summary of the most important points that are necessitated to be addressed to find consistent, reliable, reproducible and transparent QSAR models (OECD Document, 2007).

Successful development of *in silico* models to predict the carcinogenic potency of structurally diverse chemicals has been addressed in various scientific reports. A support vector machine (SVM) model was developed by Massarelli et al. (2009) using 55 chemicals with hepatocarcinogenic potency to predict the unwanted property for new chemical entities. Global and robust QSAR models were established by Kar and Roy (2011) using carcinogenicity data of 1464 structurally diverse compounds. Further, Kar and Roy (2012) developed interspecies carcinogenicity correlation models for rat and mouse based on the carcinogenic potential of 166 organic chemicals. In the above mentioned reports, the carcinogenic potency used by the authors to develop QSAR models was measured in TD_{50} scale. As discussed earlier, it is quite clear that potency measurement in OSF scale is more effective than TD_{50} scale and it is advantageous in the context of low dose-response and target organ-specific dose-response. Unfortunately, there is a lack of animal or human studies in the literature to determine OSFs. Only a single QSAR model is found using OSF values after thorough searching of scientific publications. Wang et. al. (2011) developed a QSAR model to predict the OSFs of 70 chemicals based on male/female human, rat, and mouse bioassay data obtained from the US EPA's IRIS database (http://www.epa.gov/iris/). Only internal validation was performed for this study and OSF values of 5 chemicals were wrongly reported (Wang et al., 2011) as identified in comparison with the original US EPA's IRIS database. In these perspectives, we have developed a rodent quantitative carcinogenicity model using 70 chemicals (previous 68 molecules along with 2 new molecules included in IRIS database) using the exact OSF values given in the IRIS database. Linear Discriminant Analysis (LDA) has also been applied to identify the discriminatory features between higher and lower carcinogenic compounds. The structural fragments identified by the QSAR model to be responsible for carcinogenic potency are compared with the results of LDA and Pharmacological Distribution Diagram (PDD) approaches. The present work is aimed at determining an initial carcinogenicity classification of diverse chemicals so that they can be predicted as more or less toxic at the initial stage and finally development of a statistical regression model to derive specific information regarding the contribution of different structural and physicochemical components towards carcinogenicity.

2. Materials and methods

2.1. Dataset

The OSF of 72 chemicals experimented on rat, mouse and human expressed as mg/kg/day was reported in the USEPA's IRIS database (http://www.epa.gov/iris/). To develop rodent quantitative carcinogenicity models, Benzene and Benzidine (as bioassays were performed on human) were excluded. As we are using only rodent carcinogenic potency data for development of models, it fully complies with **OECD principle 1** (Defining the endpoint). OSFs were converted to mmol-based unit instead of mg-based unit by dividing the OSFs by their respective molecular weights prior to the development of the QSARs to prevent the influence of the molecular weight of the compounds. Then, OSF values were transformed into negative logarithmic function, thus obtaining corresponding pOSF indices. Hence high value of pOSF means high carcinogenic potency.

2.2. Descriptor calculation and dataset splitting

A pool of 447 descriptors was calculated using Dragon 6 (TALETE srl, Italy), Cerius 2 version 4.10 (Accelrys Inc., San Diego, CA) and Hyperchem Release 8.0.3 for windows (Hypercube Inc.) software. Data set splitting and methodological steps performed for LDA and QSAR, descriptors thinning for LDA technique are schematically represented in Fig. 1. Four compounds [Aniline(**C05**), Benzotrichloride **(C11)**, Bis(chloromethyl)ether (**C14)** and 4,4′ Methylenebis(N,N′-dimethyl)aniline (**C51)]** were excluded during the development of regression based QSAR models based on a preliminary analysis though all compounds were considered during the classification QSAR model development.

The normality distribution of the response values of the training set data was checked using different statistical tests. The normality distribution result and plot are presented in Fig. S1 in Supplementary Materials section.

2.3. Chemometric tools for development of classification and regression based models

To classify higher and lower carcinogenic chemicals, the LDA approach was employed. LDA is a well-known classification technique for feature extraction and dimension reduction (Mitteroecker and Bookstein, 2011). The mean value of the carcinogenic potency (pOSF) data distribution (obtained from the normality distribution plot) is 2.80 (in logarithmic scale). The corresponding dose response value of the mentioned pOSF value is 0.5 mg/kg/day which is taken as the threshold for our LDA analysis. Compounds having the OSF value of 0.5 mg/kg/day or less than 0.5 mg/kg/day, are classified as higher carcinogenic. For development of regression based QSAR models, initially stepwise regression (Darlington, 1990) was carried out and then the variables selected in stepwise regression were subjected to partial least squares (PLS) (Wold, 1995) analysis. Also, statistical techniques like genetic function approximation (GFA-MLR) followed by multiple linear regression (Fan et al., 2001) and genetic partial least squares (G/PLS) (Rogers and Hopfinger, 1994; Wold, 1995) were applied. Sections 2.2 to 2.3 are written in compliance with **OECD principle 2** (Defining the algorithm).

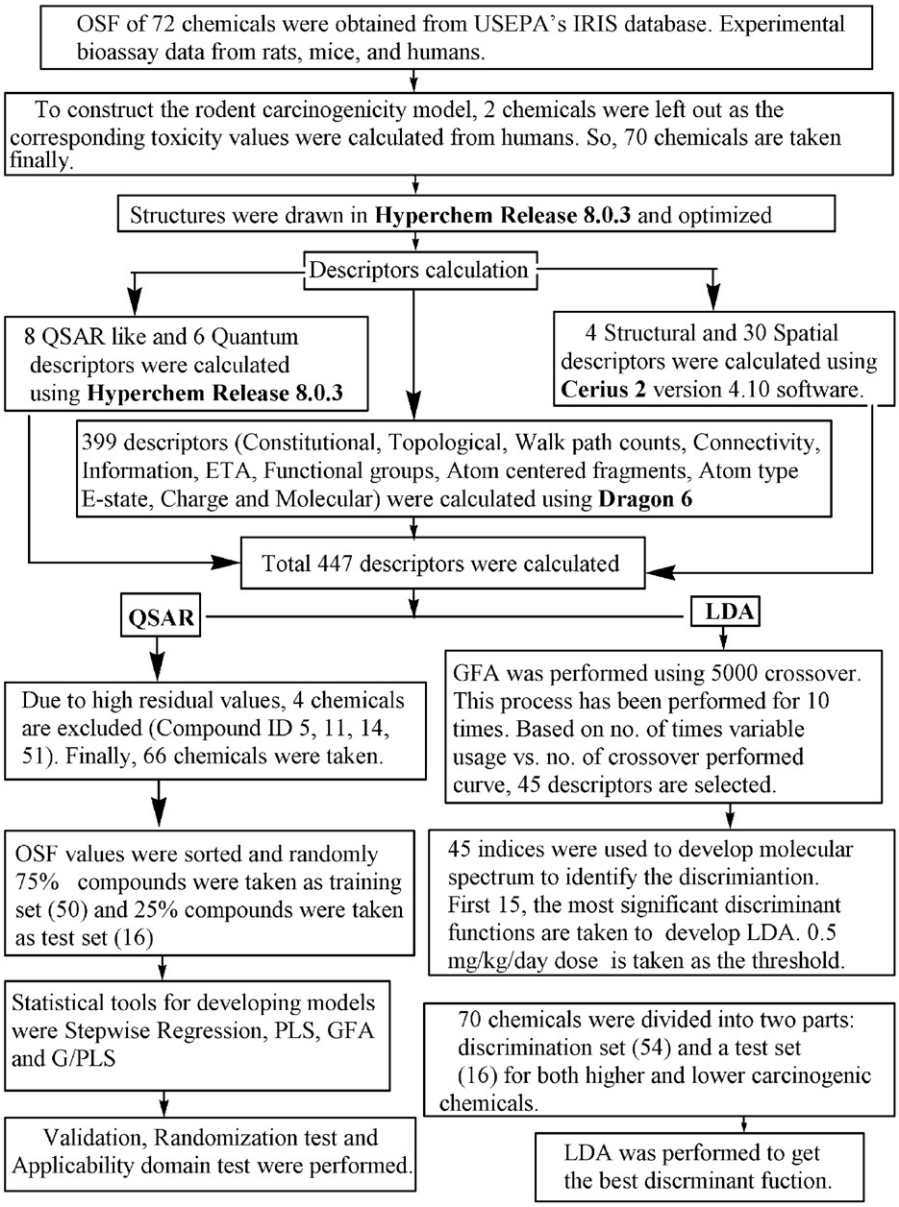


Fig. 1. Schematic diagram of methodological steps performed for regression based QSAR and LDA.

2.4. Software

Software tools like STATISTICA 7.0 (STATSOFT Inc., USA), SPSS 9.0 (SPSS Inc., USA), MINITAB 14 (Minitab Inc., USA) and SIMCA-P 10.0 (UMETRICS, Umea, Sweden, 2002) have been used in the present study.

2.5. Validation metrics for classification QSAR models

In the classification technique, validation may be performed to assess the performance of the model in terms of correct qualitative prediction of the dependent variable. Most commonly used validation parameters in classification techniques are Accuracy, Sensitivity, Specificity, Precision, F-measure (Roy and Mitra, 2011). To evaluate the classifier model performance and classification capability, a number of statistical tests have been employed. Such tests include computation of Wilk's λ statistics (Gálvez-Llompart et al., 2011), Canonical index (R_c) (Prado-Prado et al., 2009), Matthews correlation coefficient (MCC) (Matthews, 1975), squared Mahalanobis distance (Gálvez-Llompart et al., 2011) and plotting of Receiver Operating Characteristic (ROC) curve (Fawcett, 2006). The method used to select the descriptors was based on the Fisher-Snedecor parameter (F), which determines the relative importance of candidate variables (Gálvez-Llompart et al., 2011). Another statistic which was used in the model was the chi-square χ^2 to test the independence between groups. We also took into consideration the probability-level ($p < 0.05$) and the proportion between the cases and variables (ρ).

The calculation of the said statistical parameters involves determination of the degree of correctiveness of predicted classification of compounds with respect to their assigned *a priori* class. The ROC curve identifies the discrimination ability of the classification system and the graph is obtained by plotting the sensitivity and (1-specificity) indices along the Y and X axes respectively. The performance of a diagnostic variable can be quantified by calculating the area under the ROC curve (AUROC). The ideal test would have an AUROC of 1, whereas a random guess would have an AUROC of 0.5 (Fawcett, 2006). In this study we have calculated two new additional parameters (Perez-Garrido et al., 2011) namely the ROC graph Euclidean distance (ROCED) and the ROC graph Euclidean distance corrected with Fitness Function (FIT(λ)) (ROCFIT) to have better explainable results. Mathematical formulae of the parameters along with the theory of ROC graph have been discussed in the Supplementary Materials section. PDD (Murcia-Soler et al., 2003) is a graphical representation that provides a straightforward way of visualizing the regions of minimum overlap between active and inactive compounds or higher toxic and lower toxic compounds, as well as the regions in which the probability of finding active compounds is maximum.

2.6. Validation metrics for regression based QSAR models

The robustness of the regression models was verified by using different types of validation criteria. **Section 2.6** is written in compliance with **OECD principle 4** (Defining goodness-of-fit and robustness and defining predictivity). The quality of

the equations was judged by the quality metric R^2, as well as the following internal validation metrics: the leave-one-out cross validation parameter Q^2_{LOO}, leave-many-out (LMO) cross-validation metrics (in this study, we have used L-10%-O, L-25%-O and L-50%-O for the best model) (Geisser, 1975), external validation metrics: R^2_{pred} or $Q^2_{ext(F1)}$. The r^2_m metrics namely $\overline{r^2_m}$ and Δr^2_m (Ojha et al., 2011) have been developed by the present authors' group for internal, external and overall validation of models. It has been shown that the value of $\Delta r^2_{m(test)}$ should preferably be lower than 0.2 provided that the value of $\overline{r^2_{m(test)}}$ is more than 0.5 (Roy et al., 2012). Similarly, $\overline{r^2_{m(LOO)}}$ and $\Delta r^2_{m(LOO)}$ parameters can be used for the training set and $\overline{r^2_{m(overall)}}$ and $\Delta r^2_{m(overall)}$ can be used for the overall set (Roy et al., 2012). The r^2_m metrics have been used widely by our group (Kar et al., 2010) as well as different other groups (Toropova et al., 2010; Shahlaei et al., 2010) to check predictive ability of the developed QSAR models. More details about the r^2_m metrics are available at https://sites.google.com/site/rm2forqsarvalidation/.

The models were also subjected to the test for criteria of external validation as suggested by Golbraikh and Tropsha (2002). Additional validation parameters like $Q^2_{ext(F2)}$ (Schüürmann et al., 2008) and ($Q^2_{ext(F3)}$) (Consonni et al., 2010) have also tried to check the model reliability.

2.7. Y-randomization for regression based QSAR models

The best QSAR model (PLS model) was also subjected to a randomization test (Wold et al., 1998). In an ideal case, the average of R^2s for the randomized models should be zero, i.e. R^2_r should be zero. Accordingly, we have calculated the metric $^cR^2_p$ using the following formula (Mitra et al., 2010):

$$^cR^2_p = R \times \sqrt{R^2 - R^2_r} \quad (1)$$

For an acceptable model, the value of $^cR^2_p$ should be more than 0.5.

2.8. Applicability domain (AD) and limits of applicability of regression based QSAR model

According to **OECD principle 3**, a QSAR model should be reported with a defined domain of applicability. Technically, AD represent the chemical space defined by the structural information of the chemicals used in model development, i.e., the training set compounds in a QSAR analysis.Here, we have tried two different approaches to assess AD. The applicability domain of the models was checked using the (a) leverage approach (Gramatica, 2007) and (b) the DModX (distance to the model in X-space) approach (Wold et al., 2001).

3. Result and discussions

3.1. Results obtained from the discriminant analysis

Considering the threshold value of 0.5 mg/kg/day (as stated earlier), out of 70 compounds, 23 are identified as lower carcinogenic compounds and 47 are identified as higher carcinogenic compounds. To perform discrimination analysis, initially the total pool of 447 descriptors was selected. As the number of descriptors is too high for the LDA approach, a descriptor thinning approach was used and finally 45 descriptors are selected. The thinning approach is thoroughly discussed in Fig. 1.Then, molecular spectrum (Murcia-Soler et al., 2003) was designed using 45 indices corresponding to the total 70 compounds under study, and is presented in Fig. S2 in Supplementary Materials section. The molecular spectrum shows different profiles of the higher carcinogenic and lower carcinogenic groups, with clearly differentiated zones for certain indices. Analyzing molecular spectrum, finally 15 indices were taken into account for LDA model development.

The set of 70 molecules was divided into two parts: a discrimination set and a test set including both the higher and lower carcinogenic molecules. Out of the 54 discrimination set compounds, 35 compounds belong to the higher carcinogenic group and 19 compounds belong to the lower carcinogenic group. For the test set, 12 compounds belong to the higher carcinogenic and 4 compounds belong to the lower carcinogenic group.

LDA was performed setting forward stepwise method of variable selection with F=4 for inclusion; F=3.9 for exclusion. The priori classification probabilities are set to same (0.5) for all groups. The best discrimination function was obtained with the variables *nRNNOx*, *Cl−086*, *MAXDP* and *Wap*. The discriminant function ΔP is represented with the following equation:

$$\begin{aligned}
&\Delta P = 6.51 \times nRNNOx + 6.84 \times Cl{-}086 \\
&\quad + 8.58 \times Wap + 3.59 \times MAXDP - 3.489 \\
&n_{Tr} = 54, \lambda = 0.483, R_c = 0.719, Mahalanobis_distance = 4.522, \\
&MCC_{Training} = 0.714, AUROC_{Training} = 0.898 \\
&F(df = 4,49) = 13.118; (p < 0.0000), \\
&\chi^2(df = 12) = 36.40; (p < 0.0000); \rho = 13.5; . \\
&n_{Test} = 16, MCC_{Test} = 0.667, AUROC_{Test} = 0.896 \\
&ROCED = 0.773, ROCFIT = 1.6
\end{aligned} \quad (2)$$

The LDA equation is comprised of only four independent variables. The statistical data and parameters strongly account for the significance of this derived equation. All the metrics are within the acceptable limit for a reliable and acceptable LDA model. The model correctly classified 34 compounds out of 35 compounds as higher carcinogenic compounds and 13 compounds out of 19 compounds as lower carcinogenic compounds for the discrimination set. The discrimination set showed the following results: sensitivity=97.1%, specificity= 68.4%, precision=85%, accuracy=87.0% and F-measure=90.7%. The developed LDA model was later used to predict the test set to validate the model externally. The LDA model correctly classified 11 compounds out of 12 compounds as higher carcinogenic compounds and 3 compounds out of 4 compounds as lower carcinogenic compounds for the test set. The obtained validation parameters for the test set are also encouraging. The results are as follows: sensitivity=91.7%, specificity=75%, precision=91.7%, accuracy=87.5% and F-measure=91.7%. All the results are obtained based on the classification matrix obtained by LDA.

The area under the ROC curve (AUROC) was also determined to check the performance of the classification model for both the discrimination and test sets. The calculated values of AUROC for discrimination and test set are 0.898 and 0.896 respectively. The results are quite on the higher side of the acceptable limit of 0.5. AUROC also strongly supports the reliability of our developed discrimination model. ROC curves for the discrimination and test sets are represented in Fig. S3 in Supplementary Materials section. The ROCED parameter can take values between 0 (perfect classifier for both training and test set) to 4.5 (random classifier). The parameter ROCED bears a value of 0 for a perfect classifier, a value greater than 2.5 is considered as random classifier and above 4 is considered as bad classifier. Our model showed a value of 0.773 for ROCED, which corresponds to a good quality of the ROC analysis. ROCFIT was calculated by dividing the ROCED with Wilk's λ value, and an acceptable value of 1.6 was obtained. These two parameters prove the following points: 1. the obtained model has a similar accuracy for the training and test series, 2. both training and test sets have ratings close to perfection and 3. A maximum accuracy on the test set. The MCC usually varies from −1 to +1 referring to an inverse classification to a perfect classification respectively, whereas a value of 0 corresponds to random classification performance. The present study also showed an acceptable value for the MCC; a value of 0.714 for the training and 0.667 for the test set. The training set shows a near perfect classifier value of MCC (0.714) than the test set, where the value is still close to perfect classification instance.

The discriminant equation has been used to calculate the DF value for all the compounds from which PDD was developed for the discrimination and test set compounds. Table 1 shows the DF (ΔP) of each compound for discrimination and test sets, obtained as the difference between the variables defining the groups of higher carcinogenic and lower carcinogenic molecules. Molecules with DF values higher than 0.5 ($\Delta P > 0.5$) were classified as lower carcinogenic, while $\Delta P < 0$ corresponds to higher carcinogenic

Table 1
Results obtained by linear discriminant analysis and regression analysis carried out with 70 compounds. Observed and calculated/predicted OSF carcinogenicities of 66 chemicals for which quantitative models are constructed.

CompoundID	Chemical name	*In silico* methods					
		LDA				Regression based QSAR	
		OSF mg/kg/day	Classification based on threshold value 0.5mg/kg/day	Discriminant Function (ΔP)	Classification with developed LDA model	Observed carcinogenicity Log_{10}(MW/OSF)	Calculated/predicted carcinogenicity Log_{10}(MW/OSF) (PLS Model)
Discrimination set							
C02	Acrylamide	0.5	P	−1.891	+	2.153	3.229
C03	Acrylonitrile	0.54	P	−2.514	+	1.992	2.825
C04	Aldrin	17	N	2.756	−	1.332	2.064
C05	Aniline	0.0057	P	−2.581	+	NU	NU
C10	Benzo[a]pyrene	7.3	N	5.290	−	1.539	2.285
C11	Benzotrichloride	13	N	−2.462	+	NU	NU
C12	Benzyl chloride	0.17	P	−1.390	+	2.872	2.879
C13	Bis(chloroethyl)ether	1.1	N	−0.313	+	2.114	2.699
C14	Bis(chloromethyl)ether	220	N	−2.914	+	NU	NU
C15	Bromate	0.7	P	−2.434	+	2.265	2.937
C16	Bromodichloromethane	0.062	P	−2.941	+	3.422	3.467
C18	Carbon tetrachloride	0.07	P	−3.024	+	3.342	3.561
C19	Chlordane	0.35	P	3.934	−	3.068	1.657
C20	Di(2-ethylhexyl)adipate	0.0012	P	0.194	U	5.490	4.653
C21	Di(2-ethylhexyl)phthalate	0.014	P	1.379	−	4.446	5.197
C22	Dibromochloromethane	0.084	P	−2.844	+	3.394	3.800
C23	1,2-Dibromoethane	2	N	−3.196	+	1.973	2.630
C24	3,3′-Dichlorobenzidine	0.45	P	−1.807	+	2.750	2.709
C25	p,p′–DDA	0.24	P	−1.677	+	3.125	2.947
C26	p,p′–DDE	0.34	P	−1.755	+	2.971	2.950
C27	DDT	0.34	P	−1.621	+	3.018	3.035
C29	Dichloromethane	0.0075	P	−3.067	+	4.054	2.694
C30	1,3-Dichloropropene	0.05	P	−1.687	+	3.346	2.255
C31	Dichlorvos	0.29	P	−0.851	+	2.882	2.896
C32	Dieldrin	16	N	4.768	−	1.377	2.064
C33	1,4-Dioxane	0.1	P	−2.535	+	2.945	2.513
C35	Epichlorohydrin	0.0099	P	−1.550	+	3.971	2.598
C36	Folpet	0.0035	P	0.133	U	4.928	5.341
C38	Furmecyclox	0.03	P	0.532	−	3.923	4.330
C40	Heptachlor epoxide	9.1	N	3.602	−	1.631	1.881
C41	Hexachlorobenzene	1.6	N	−2.363	+	2.250	2.433
C42	Hexachlorobutadiene	0.078	P	−2.616	+	3.524	2.818
C43	alpha-HCH	6.3	N	4.613	−	1.664	2.067
C45	Technical HCH	1.8	N	4.613	−	2.208	2.067
C46	Hexachlorodibenzo-p-dioxin	6200	N	−0.036	+	−1.200	-0.858
C49	Hydrazine/Hydrazine sulfate	3	N	−3.489	+	1.029	1.353
C50	Isophorone	0.00095	P	−0.823	+	5.163	4.237
C51	4,4′-Methylenebis(N,N′-dimethyl)aniline	0.046	P	−2.559	+	NU	NU
C52	N-Nitroso-di-n-butylamine	5.4	N	5.126	−	1.467	1.503
C53	N-Nitroso-N-methylethylamine	22	N	4.583	−	0.603	0.169
C54	N-Nitrosodi-N-propylamine	7	N	4.955	−	1.270	1.348
C55	N-Nitrosodiethanolamine	2.8	N	4.819	−	1.680	1.163
C56	N-Nitrosodiethylamine	150	N	4.736	−	−0.167	1.132
C58	N-Nitrosodiphenylamine	0.0049	P	−0.7	+	4.607	4.295
C59	N-Nitrosopyrrolidine	2.1	N	4.780	−	1.678	1.213
C60	Pentachlorophenol	0.4	P	−1.303	+	2.823	3.426
C61	Prochloraz	0.15	P	0.891	−	3.400	2.914
C62	Propylene oxide	0.24	P	−2.705	+	2.384	2.379
C65	1,1,2,2-Tetrachloroethane	0.2	P	−2.835	+	2.924	3.193
C66	1,1,2-Trichloroethane	0.057	P	−1.710	+	3.369	2.923
C67	2,4,6-Trichlorophenol	0.011	P	−1.471	+	4.254	3.480
C69	2,4,6-Trinitrotoluene	0.03	P	−0.955	+	3.879	4.001
C70	Vinyl chloride	0.72	P	−3.068	+	1.939	1.908

Table 1 (*continued*)

CompoundID	Chemical name	*In silico* methods					
		LDA				Regression based QSAR	
		OSF mg/kg/day	Classification based on threshold value 0.5mg/kg/day	Discriminant Function (ΔP)	Classification with developed LDA model	Observed carcinogenicity Log_{10}(MW/OSF)	Calculated/predicted carcinogenicity Log_{10}(MW/OSF) (PLS Model)
C71	BDE-209	0.0007	P	−0.553	+	6.137	5.933
Test set							
C01	Acephate	0.0087	P	−0.771	+	4.323	2.870
C06	Aramite	0.025	P	1.014	−	4.127	2.878
C07	Azobenzene	0.11	P	−2.430	+	3.219	2.643
C17	Bromoform	0.0079	P	−3.259	+	4.505	3.637
C28	1,2-Dichloroethane	0.091	P	−0.599	+	3.036	2.639
C34	1,2-Diphenylhydrazine	0.8	P	−2.746	+	2.362	2.274
C37	Fomesafen	0.19	P	1.822	−	3.364	2.383
C39	Heptachlor	4.5	N	2.587	−	1.919	1.883
C44	beta-HCH	1.8	N	4.614	−	2.208	2.067
C47	Hexachloroethane	0.014	P	−2.808	+	4.228	3.731
C48	Hexahydro-1,3,5-trinitro-1,3,5-triazine	0.11	P	−1.125	+	3.305	3.670
C57	N-Nitrosodimethylamine	51	N	4.431	−	0.162	−0.804
C63	Quinoline	3	N	−2.54	+	1.634	2.722
C64	1,1,1,2-Tetrachloroethane	0.026	P	−1.694	+	3.810	3.194
C68	Trifluralin	0.0077	P	0.269	U	4.639	4.741
C72	Dichloroacetic acid	0.05	P	−1.913	+	3.411	3.616

P=Higher carcinogenic compounds (bioassay value 0.5 mg/kg/day or less than that), *N*=Lower carcinogenic compounds (bioassay value more than 0.5 mg/kg/day). Discriminant function values higher than 0.5 ($\Delta P > 0.5$) were classified as lower carcinogenic which is assigned with (−), while $\Delta P < 0$ corresponds to higher carcinogenic molecules which is assigned with (+), and the compounds with ΔP values between 0 and 0.5 were classified as undetermined carcinogenicity which is assigned with (U). NU-Not used in QSAR model development.

Discrimination set:

For the higher carcinogenic group:

Undetermined (U)=5.714%, False prediction (−)=11.429%, Overall accuracy=82.857%, Adjusted accuracy=87.879%.

For the lower carcinogenic group:

Undetermined (U)=0%, False prediction (+)=36.842%, Overall accuracy=63.158%, Adjusted accuracy=63.158%.

Test set:

For the higher carcinogenic group:

Undetermined (U)=8.333%, False prediction (−)=16.667%, Overall accuracy=75%, Adjusted accuracy=81.818%.

For the lower carcinogenic group:

Undetermined (U)=0%, False prediction (+)=25%, Overall accuracy=75%, Adjusted accuracy=75%.

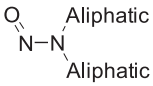
molecules, and the compounds with ΔP values between 0 and 0.5 were classified as molecules with undetermined carcinogenicity. These discriminant conditions are imposed to minimize the percentage of error, i.e. to give the lowest possible number of false positives.

The overall accuracy was 82.9% in the higher carcinogenic group and 63.2% in the lower carcinogenic group for the discrimination set. The cross-validation test was applied to the ΔP function with a group of 4 lower carcinogenic and 12 higher carcinogenic molecules not used in the discriminant function. The fact is that both the overall accuracy (75% for both higher carcinogenic and lower carcinogenic group) and the adjusted accuracy (81.8% for higher carcinogenic group, 75% for lower carcinogenic group) for the test set are quite similar to the discrimination set.

The discrimination of carcinogenicity was carried out to show that the obtained discriminant function values in the LDA for both the groups make it possible to separate the two populations. To design the PDD, we observed that the maximum of the E_i (expectancy to get lower carcinogenic compounds) and E_a (expectancy to get higher carcinogenic compounds) values are distributed on different sides of $\Delta P=0$. We obtained positive values for lower carcinogenic compounds (with a maximum value around $\Delta P=6$) and negative values for higher carcinogenic compounds (with a maximum value of $\Delta P=-4$, approximately), in both the discrimination and the test groups. The values calculated for the discriminant function and the corresponding classification appear in Table 1. PDDs are presented in Fig. S4 for the discrimination set and the test set, respectively in Supplementary materials section. On analysis of Fig. S4, although overlapping of E_i can be seen in the E_a region for the discrimination set of compounds, the overlapping building blocks are significantly lower. On the other side, only one block of E_i is overlapped in the E_a region for the test set. Less is the overlapping between E_i and E_a, more meaningful is the PDD. The studied PDD for the test set is more significant and reliable than the discrimination set as the overlapping of E_i and E_a are significantly less. Using the PDD, it is possible to discriminate between higher and lower carcinogenic groups within a structurally heterogeneous set of compounds and it constitutes a valuable tool in the validation of discrimination analysis for our study.

A contribution plot (Fig. 2) was developed for the best discriminating descriptors (Eq. (2)) by taking the product of their average values with their corresponding coefficients as in the discriminant Equation. The indices *nRNNOx*, *Cl−086* and *Wap* show marked positive contributions to the lower carcinogenic group. This demonstrates the importance of the presence of

number of aliphatic N-nitroso groups



(*nRNNOx* index) and hydrophobicity measure of Cl atom attached to sp_3 hybridized carbon (C1) atom (*Cl*−086 index, an atom centered fragments) in relation to the lower carcinogenicity of our studied compounds. *WAP* is the Wiener index, which is the sum of the number of edges in the shortest paths in a chemical graph between all pairs of non-hydrogen atoms in a molecule. Fig. 2 confirms that these three indices are the major discriminatory features between higher and lower carcinogenic groups. Again, if we compare among these three indices, *nRNNOx* distinctly makes the major difference between two groups. It is quite significant that all the compounds comprising of *nRNNOx* group are identified as the lower carcinogenic compound. It is quite interesting to point out that N-Nitrosodiphenylamine (**C58**) which contains aromatic N-nitroso fragment shows higher carcinogenicity. So, it is quite clear that though nitroso fragment containing compounds are carcinogenic but aromatic N-nitroso compounds are higher carcinogens than the aliphatic N-nitroso compounds. In our discrimination analysis studies only aliphatic N-nitroso containing compounds are showing lower carcinogenic property very clearly. Further explanation on the mechanistic contribution of this feature to the carcinogenicity is given in the QSAR interpretation section in details.

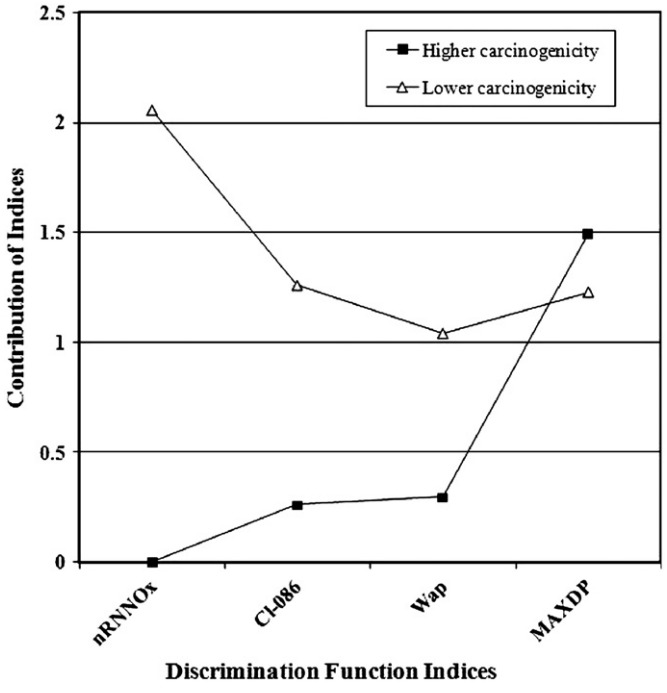


Fig. 2. Average contribution of indices to the discriminant functions for higher and lower carcinogenic molecules groups.

Though *Cl*−086 and *WAP* are not as significant as *nRNNOx*, but effect of *Cl*−086 is also observed in the discrimination between higher and lower carcinogenic compounds. Compounds like alpha-HCH (**C43**), Heptachlor epoxide (**C40**), Aldrin (**C04**) are classified as lower carcinogenic compounds as they contain high number of the *Cl*−086 feature (6, 3 and 2, respectively). On the contrary, though Chlordane (**C19**) contains 4 *Cl*−086 feature but it is classified in the higher carcinogenic group. The other fragmental properties and physicochemical properties have effective contribution to higher carcinogenicity of Chlordane.

Significantly, MAXDP has a positive contribution towards higher carcinogenic compounds. MAXDP is a topological index that represents the maximum positive intrinsic state difference and can be related to the electrophilicity of the molecule (Kier et al., 1991). The result of contribution plot signifies that compounds carrying electrophilic property shows higher carcinogenic value and vice versa. Further analysis has been performed in the regression analysis part.

3.2. Results obtained from the regression analysis

Statistically significant QSAR models were developed using different chemometric tools. A detailed report of the statistical quality of various models is elaborated in Table 2. Among the various models developed, the best results were obtained from the PLS technique based on giving equal importance to internal as well as external predictivity. The developed equation is as follows:

$$\begin{aligned}\log_{10}(\mathrm{MW/OSF}) &= 0.237 - 2.112 \times (\mathrm{nRNNOx}) \\ &\quad + 0.626 \times (\mathrm{MAXDP}) - 0.499 \times (\mathrm{Cl-089}) \\ &\quad + 2.415 \times (\mathrm{Mp}) - 0.231 \times (\mathrm{nCXr}) - 1.633 \times (\mathrm{nArOR}) \\ &\quad + 0.252 \times (\mathrm{nArX}) - 0.804 \times (C-005) \\ n_{Training} &= 50, LV = 7, R^2 = 0.800, Q^2_{LOO} = 0.643, \\ Q^2_{L-10\%-0} &= 0.716, Q^2_{L-25\%-O} = 0.711, Q^2_{L-50\%-O} = 0.717 \\ n_{Test} &= 16, Q^2_{F1} = R^2_{pred} = 0.645, \overline{r^2_{m(test)}} = 0.624, \Delta r^2_{m(test)} = 0.155, \\ Q^2_{F2} &= 0.605, Q^2_{F3} = 0.714 \end{aligned} \quad (3)$$

Eq. (3) involving 8 descriptors and 7 latent variables (LVs) could explain 80.0% of the variance. The statistical significance of the developed model is reflected from the acceptable values of Q^2_{LOO} (0.643) and $Q^2_{\mathrm{ext(F1)}}$ (0.645). To check the model reliability in terms of internal validation, the leave-many-out (LMO) cross-validation (10%, 25% and 50%) tools were also applied. LMO cross-validation for 10%, 25% and 50% data point removal could predict 71.6%, 71.1% and 71.7% respectively of total variance. The compounds deleted in different cycles of leave-many-out cross-validation are indicated in Table S1 in Supplementary Materials section. The obtained value of $\overline{r^2_{m(test)}}$ is more than 0.5 and the value of differences between r^2_m and r'^2_m metrics ($\Delta r^2_{m(test)}$) is less than 0.2 inferring that the predicted OSF carcinogenicity value for these molecules, calculated using the above equation, are in close proximity to the experimental data.

Quite close values for the $Q^2_{\mathrm{ext(F1)}}$ (0.645) and $Q^2_{\mathrm{ext(F2)}}$ (0.605) parameters indicate that the test set selected for the QSAR model

Table 2
Statistical quality of the developed carcinogenicity QSAR models.

Response variable	Model No.	Statistical Tool	No. of Descriptors	R^2	Q^2	R^2_{pred}	$\overline{r^2_{m(test)}}$	$\Delta r^2_{m(test)}$
Carcinogenicity	1	Stepwise MLR	9	0.822	0.706	0.636	0.621	0.204
	2[a]	**PLS**	**8 (LV=7)**	**0.800**	**0.643**	**0.645**	**0.624**	**0.155**
	3	GFA (Linear)	6	0.711	0.599	0.572	0.471	0.231
	4	G/PLS (Linear)	7 (LV=4)	0.683	0.547	0.730	0.652	0.086

[a] The best model.

development has similar distribution of response as the training set. Thus, the model may be considered statistically significant and satisfactory for predicting the carcinogenicity of a new set of molecules. Again, the function $Q^2_{ext(F3)}$ (0.714) which is independent of the external data distribution and fulfils some basic mathematical properties such as ergodic and associative properties, also shows a satisfactory value for the model. Acceptable values of all these parameters for the best model indicate that the obtained model has statistical reliability and good internal as well as external predictive potential. Moreover, the carcinogenicity values of all the compounds calculated/predicted using Eq. (3) was plotted against the observed carcinogenicity data and the resulting graph (Fig. S5 in Supplementary Materials section) showed that the points were limitedly scattered about the line of fit. This again implicated the predictive efficacy of the developed QSAR model. The PLS model also satisfied the statistical validation parameters set forth by Golbraikh and Tropsha (2002). For the PLS model, these statistical parameters yielded the following results: $Q^2=0.643$, $r^2=0.727$, $(r^2-r_0^2)/r^2=0.085$, $(r^2-r_0'^2)/r^2=0.002$, $k=0.86$, $k'=0.85$.

The VIPs (Variable Importance Projections) and the coefficient histogram of the original descriptors for Model 2 are presented as histograms in Fig. S6 and Fig. S7 in Supplementary Materials section. To comply with the **OECD Principle 5**, mechanistic interpretation should be given for any predictive QSAR model as far as possible. Here, we explain the interpretation and importance of each descriptor appearing in the best regression equation modeling rodent carcinogenicity with suitable examples:

(a) *nRNNOx*is the most important descriptor according to the VIP plot and it has a negative contribution towards carcinogenicity. It signifies the number of aliphatic N-nitroso groups. As the descriptor has a negative contribution towards the carcinogenicity, presence of *nRNNOx* functional group decreases the carcinogenicity, and with the absence of *nRNNOx*, carcinogenicity of a compound increases. N-Nitroso-N-methylethylamine (**C53**), N-Nitrosodi-N-propylamine (**C54**) and N-Nitrosodiethylamine (**C56**) comprising of aliphatic N-nitroso group showing lower carcinogenicity values (0.603, 1.270 and -0.167, respectively). On the contrary, compounds like Di(2-ethylhexyl)adipate (**C20**), Isophorone (**C50**) and BDE-209 (**C71**) are showing high carcinogenic value due to the absence of *nRNNOx*. Quite interestingly, though N-Nitrosodiphenylamine (**C58**) is a nitroso compound but it contains aromatic N-nitroso fragment; as a result it shows high carcinogenicity value (4.607). Again, it is worth mentioning that *nRNNOx* fragment is also the most important discriminating index between higher and lower carcinogenic property. So, further explanation has been given below regarding the mechanism of this particular fragment towards carcinogenicity.
Major N-nitroso compounds have been classified by the IARC as Group 2B carcinogens (IARC Monographs, 2006). The cytotoxic effect of major N-nitroso compounds on tumor cell is known to result from the DNA binding of alkylating species generated during metabolic composition. The DNA adduct is mainly formed by S_N2 nitrosation mechanism (Enoch and Cronin, 2010). The reaction is presented in Fig. S8 in Supplementary Materials section. Another important pathway is confirmed by Tanno et al. (1996) that the generation of NO from aromatic N-nitroso compounds at ambient temperature by trapping of NO as a nitrosyl complex. Thus, the cytotoxicity originating from NO is expected to be potent in aromatic N-nitroso compounds as they have NO-generating ability, compared with that in aliphatic N-nitroso compounds. A conjugating effect between the aromatic ring carbon and neighboring nitrogen influences the NO generating ability for aromatic N-nitroso compounds. On the other hand, NO production from the aliphatic N-nitroso compounds was not observed and those N-nitroso compounds did not show effective cytotoxic activity. The above mentioned mechanism and our developed QSAR equation can fully support why aromatic N-nitoso compounds are showing higher carcinogenic property than the aliphatic N-nitroso compounds.

(b) MAXDP is the maximum positive intrinsic state difference and it can be related to the electrophilicity of the molecule. Electrophilicity contributes postively towards carcinogenicity and it acts as one of the most important features in carcinogenicity prediction for our studied compounds. Compounds carrying electrophilic property shows higher carcinogenic values and vice versa. Most genotoxic carcinogens are strong electrophiles. Chemical carcinogens are converted into electrophilic reactants and/or metabolites with electron-deficient sites. These electrophilic compounds can then exert their carcinogenic effects through covalent interaction with cellular macromolecules (Klaassen and Watkins, 1999). To identify specific structural alerts of carcinogenicity, Ashby and Tennant (Ashby and Tennant, 1991; Putz et al., 2011) found that the majority of the rodent carcinogens were among the group of chemicals containing an electrophilic alert. All the previous studies strongly support our obtained interpretation.

(c) *Cl*−089, an atom centered fragment descriptor, refers to the hydrophobicity measure of Cl atom attached to a sp_2 hybridized carbon (C1) atom. *Cl*−089 has a detrimental effect to carcinogenicity according to our model. Hexachlorodibenzo-p-dioxin (**C46**), Hexachlorobenzene (**C41**) and Aldrin (**C04**) show lower carcinogenic profile since they contain high number of *Cl*−089 fragments (6, 6 and 2, respectively).

(d) *Mp* is the mean atomic polarizability which is a constitutional descriptor, and is measured by summation of the atomic contributions (Hemmateenejad et al., 2005). Atomic polarizability is a sum over all atoms in the molecule and describes the molecule's ability to polarize in a magnetic field. The more polarizable molecules are more carcinogenic according to our developed QSAR model due to the positive contribution towards carcinogenic property. Increase in carcinogenicity with atomic polarizability is already reported in the literature (Jelcic, 2004).

(e) nCXr is defined as the number of X (halogen) on ring C(sp3). This functional group count descriptor has a negative contribution towards carcinogenicity. Aldrin (**C04**) and Dieldrin (**C32**) are showing lower carcinogenicity values due to presence of high number of chlorine atoms on ring C(sp3).

(f) nArOR can be explained as the number of ether linkages (aromatic) which has a detrimental effect towards carcinogenicity. Hexachlorodibenzo-p-dioxin (**C46**) contains 2 ether linkages between aromatic groups and shows low carcinogenicity value.

(g) nArX is defined as the number of X (halogen) on an aromatic ring. This functional group count descriptor is conducive to carcinogenicity. Compounds like BDE-209 (**C71**) is showing high carcinogenicity value due to high number of nArX count (10). Again, Hexachlorodibenzo-p-dioxin (**C46**) is showing a very poor carcinogenicity value though it contains 6 nArX functional counts. In this particular case, compound **C46** has high number of (6) *Cl*−089 fragment (one of the most important descriptor after *nRNNOx* and MAXDP) and 2 nArOR functional counts which are detrimental for carcinogenicity. Note that Cl-089 has a negative coefficient while nArX has a positive coefficient which indicates that bromine has more positive impact on carcinogenicity than chlorine (the present data set has no fluorine or iodine containing compounds) and

thus Cl-089 acts as a penalty factor for the nArX term in case of chlorine containing compounds.

(h) Atom centered fragment descriptor, *C*−005 is calculated based on fragment based approach for log P prediction. *C*−005 refers to the hydrophobicity measure of CH_3X fragment, where, X- represents any heteroatom (O, N, S, P, Se and halogens). According to the VIP plot, it is least significant among the obtained eight descriptors. It has a negative contribution towards the carcinogenicity. N-Nitroso-N-methylethylamine (**C53**) contains one CH_3X fragment, and hence, it shows lower carcinogenic profile. Also, the contribution of *nRNNOx* fragment cannot be denied for N-Nitroso-N-methylethylamine as it contains a *nRNNOx* fragment which negatively contributes to carcinogenicity. On the contrary, though Dichlorvos (**C31**) contains two CH_3X fragments but it shows higher carcinogenic profile due to lower number of –Cl fragments than the other studied compounds. The effect of other descriptors also exerts significant contribution for this exceptional case.

The PLS loading plot for the response variable (carcinogenicity) and the descriptors included in the final model are shown in Fig. 3. The carcinogenicity is explained significantly by the first component. The loading plot shows that the first component is dominated by the electronic parameters (*MAXDP* on the positive side, and *nRNNOx* on the negative) and the second component is dominated predominantly by solubility/hydrophobicity with *Cl*−089, *C*−005 and nArOR lying on the negative side. *Mp*, nCXr and nArX share the features of both components.

Model 2 was validated using a randomization test through randomly reordering (100 permutations) response data using SIMCA 10.0 (UMETRICS, Umea, Sweden, 2002). The randomization parameter model 2 is well above the permissible limit indicating that the model is not obtained by chance. Carcinogenicity intercept values are $R^2=(0.0, 0.077)$, $Q^2=(0.0, -0.774)$. The randomization plot is presented in Fig. S9 in Supplementary Materials section. The lack of chance correlation in the PLS model is also well reflected from the value of ${}^{c}R^2_p$ (0.761) which is higher than the acceptable threshold value of 0.5. Both results suggest that the obtained model is not derived by chance.

According to the OECD guidelines (OECD Document, 2007), it is desirable to verify the applicability domain of a model using multiple approaches. Chemicals considered in the present modeling work are diverse and mostly environmental contaminants and pesticides: the main categories include a) aliphatic alcohols, amines, nitrosamines, ethers, halogenated derivatives, b) aromatic alcohols, ethers, nitrosamines, halogenated derivatives, c) cyclic ethers. The range of the dependent variable for the training set compounds is −1.2 to 6.137 while the ranges for descriptors are as following: *nRNNOx* (0 to 1), MAXDP (0 to 5.57), *Cl*−089 (0 to 6), *Mp* (0.462 to 1.295), nCXr (0-6), nArOR(0 to 2), nArX(0 to 10) and *C*−005 (0 to 2). Based on Y-response (here, rodent carcinogenicity) of the training set, all test compounds are inside of the AD. Based on X-responses, compound **C17** (Bromoform) is considered as outside of the AD marginally for the *Mp* variable. There is not a single compound outside of the AD for other 7 descriptors derived from regression models. The applicability domain for the PLS model was checked using the leverage approach. Leverage values of training compounds **C31** (Dichlorvos), **C46** (Hexachlorodibenzo-p-dioxin) and **C71** (BDE-209) being greater than the critical value of 0.54 ($h > h^*$), these compounds behave as influential observations although they are not response outliers. All 16 test set compounds were found to be within the applicability domain of the model. Further, at 99% confidence level, DModX values of all test compounds are below the critical value of 3.719. Considering leverage and DModX processes of the AD tests, we conclude that all 16 test compounds are inside of the AD and their predictions are highly reliable. Based on the three applied approaches, we can confidently predict 15 test compounds based on the developed model after rigorous tests for validation and applicability domain check. Williams plot (Fig. S10) and DModX plot (Fig. S11) are presented in Supplementary Materials section.

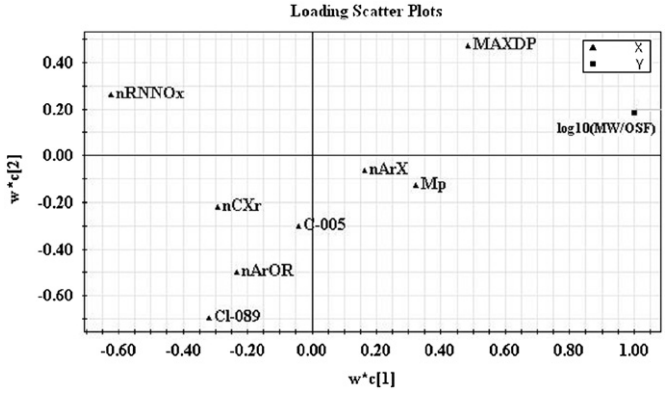


Fig. 3. The loading plot of the first two principal components for the PLS model (Model 2).

4. Comparison between interpretation of classification and regression models

The results obtained from the discriminant model and the regression models are comparable to each other. The descriptors appearing in the classification model would ideally discriminate the higher and lower carcinogenic groups. The classification model comprises of only four variables that are representative of the discriminatory features between higher and lower carcinogens. Out of four descriptors, indices *nRNNOx*, *Cl*−086 and *Wap* show marked positive contributions to the lower carcinogenic group. Making a comparison among these indices, *nRNNOx* is the most important one, which is quite clear from Fig. 2. Significantly, MAXDP

Table 3
Comparison of quality between our best models with the previous developed model.

Sl No.	Previously developed model (Wang et al. 2011)	Our developed best model
1	According to OECD principle 1, QSAR model should be developed for the defined endpoint. Here, along with the rodent (rat and mouse), human endpoint was used for 2 chemicals (Benzene and Benzidine) for the QSAR study.	Benzene and Benzidine are left out for QSAR model development and Linear Discriminatory Analysis.
2	QSAR model was developed with 70 chemicals without external validation.	External validation is performed.
3	QSAR model contains 12 latent variables. ($R^2=0.771$, $Q^2=0.732$)	Best QSAR model contains 7 latent variables. ($R^2=0.800$, $Q^2=0.643$, $R^2_{pred}=0.645$)
4	Oral slope factor values of 5 chemicals are wrongly reported if we compare with the original USEPA's IRIS database.	Models are developed after correction of all these values.

Name	Reported[a]	Original source[a]
Acrylamide	4.5	0.5
Carbon tetrachloride	0.13	0.07
1,2-Dibromoethane	85	2
1,4-Dioxane	0.011	0.1
Pentachlorophenol	0.12	0.4

[a] All values are in mg/kg/day.

has a positive contribution towards higher carcinogenic compounds. Most interestingly our QSAR results also identified *nRNNOx* fragment and MAXDP as important indices for the regression model. As the *nRNNOx* descriptor has a negative contribution towards carcinogenicity, presence of *nRNNOx* functional groups decreases the carcinogenicity and with the decreasing number of *nRNNOx*, carcinogenicity of a compound increases. Analyzing the regression model, it is quite clear that MAXDP contributes postively towards carcinogenicity (Ashby and Tennant, 1991; Putz et al., 2011). This observation is also fully supported from the results of LDA and contribution plot (Fig. 2). Structural fragments like *Cl*−089, nCXr, nArOR, nArX and *C*−005 are also identified as important features for carcinogenicity prediction. A comparison between previous work (Wang et al., 2011) and our proposed work is presented in Table 3. The major success of our studied model is that results from both techniques conclude in the same direction.

5. Conclusion

Combined analysis of the descriptors used in the best regression based QSAR equation and LDA suggest that the carcinogenicity often depends on the electrophilicity and particular structural fragments of the chemicals. Most genotoxic carcinogens are strong electrophiles. These electrophilic compounds can exert their carcinogenic effects through covalent interaction with cellular macromolecules. Based upon the descriptor MAXDP in the best QSAR model, the majority of the training set chemicals could be considered as electrophiles that interact with DNA, RNA and protein macromolecules which majorly act as nucleophiles. One of the major identified fragments for lower carcinogenic property is *nRNNOx* fragment which has a negative contribution to carcinogenicity. Most interestingly, *nRNNOx* fragment is also identified as the major discriminatory feature between higher and lower carcinogenic group by LDA analysis and it has a positive contribution towards lower carcinogenic group. It should be noteworthy to mention that all the nitroso compounds have carcinogenic property but presence of the *nRNNOx* index in a particular compound can discriminate between higher and lower carcinogens. The *Mp* descriptor signifies that more polarizable molecules are more carcinogenic. The *Cl*−089, nCXr, nArOR and *C*−005 fragments have negative contributions towards carcinogenicity. The descriptor nArX has a positive contributions towards carcinogenicity. It appears that bromine has more positive impact on carcinogenicity than chlorine (the present data set has no fluorine or iodine containing compounds) and thus Cl-089 acts as a penalty factor for the nArX term in case of chlorine containing compounds. Satisfyingly, the results obtained from regression based QSAR and LDA techniques are quite complementary to each other and the interpretations are quite similar for both techniques which help us to understand the major structural features and physicochemical properties for the carcinogenic property of our studied compounds. The structural features identified by the regression based QSAR model to be responsible for carcinogenic potency are successfully confirmed from the results of LDA and PDD approaches. Though intrinsic complexity and multistage nature of carcinogenicity is a major limitation in its prediction, the current *in silico* methods have provided some structural alerts to determine the potential carcinogenic modes of action via direct interaction with DNA and other macromolecules. The obtained results can be used as a starting point for regulatory decision making and risk assessment in future.

Acknowledgment

SK thanks the Department of Science and Technology, Government of India for awarding him a Senior Research fellowship under the INSPIRE scheme. KR thanks the Council of Scientific and Industrial Research (CSIR), New Delhi for awarding a major research project.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ecoenv.2012.05.013.

References

Ashby, J., Tennant, R.W., 1991. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemical tested by the US NTP. Mutat. Res. 257, 229–306.
Benigni, R., Giuliani, A., 2003. Putting the Predictive Toxicology Challenge into Perspective: Reflections on the Results. Bioinformatics 19, 1194–1200.
Benigni, R., Zito, R., 2004. The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results. Mutat. Res. 566, 49–63.
CDER, 1997. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for Industry. S1B. Testing for carcinogenicity of pharmaceuticals.
Cerius2, Version 4.10 Software is a Product of Accelrys Inc., San Diego, CA, USA. 〈http://www.accelrys.com/cerius2〉.
Consonni, V., Ballabio, D., Todeschini, R., 2010. Evaluation of model predictive ability by external validation techniques J. Chemometrics 24, 194–201.
Darlington, R.B., 1990. Regression and Linear models. McGrawHill, New York.
Deeb, O., 2010. Correlation ranking and stepwise regression procedures in principal components artificial neural networks modeling with application to predict toxic activity and human serum albumin binding affinity. Chemom. Intell. Lab. Syst. 104, 181–194.
DRAGON ver. 6 is software of TALETE srl, Italy, 〈http://www.talete.mi.it/products/dragon_molecular_descriptors.htm〉.
Enoch, B.S.J., Cronin, M.T.D., 2010. A review of the electrophilic reaction chemistry involved in covalent DNA binding. Crit. Rev. Toxicol. 40 (8), 728–748.
European Commission, 2006. Directive 2006/121/EC of the European Parliament and of the Council of 18 December 2006 amending Council Directive 67/548/EEC on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances in order to adapt it to Regulation (EC) No. 1907/2006 concerning the REACH and establishing a European Chemicals Agency. Official Publications of the European Communities (OPOCE), Luxembourg.
Fan, Y., Shi, L.M., Kohn, K.W., Pommier, Y., Weinstein, J.N., 2001. Quantitative Structure-Antitumor Activity Relationships of Camptothecin Analogues: Cluster Analysis and Genetic Algorithm-Based Studies. J. Med. Chem. 44, 3254–3263.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27, 861–874.
Gálvez-Llompart, M., Recio, M.C., García-Domenech, R., 2011. Topological virtual screening: a way to find new compounds active in ulcerative colitis by inhibiting NF-κB. Mol. Divers 15, 917–926.
Geisser, S., 1975. The Predictive Sample Reuse Method with Application. J. Amer. Stat. Ass. 70, 320–328.
Golbraikh, A., Tropsha, A., 2002. Beware of q2!. J. Mol. Graph. Model 20, 269–276.
Gramatica, P., 2007. Principles of QSAR Models Validation: Internal and External. QSAR Comb. Sci. 26, 694–701.
Hemmateenejad, B., Miri, R., Safarpour, M.A., Khoshneviszadeh, M., Edraki, N., 2005. Conformational analysis of some new derivatives of 4-nitroimidazolyl-1,4-dihydropyridine based calcium channel blockers. J. Mol. Struct. (Theochem.) 717, 139–152.
Hyperchem Release 8.0.3 for windows, Hypercube Inc. copyright 2007.
IARC Monographs on the evaluation of carcinogenic risks to humans, Lyon, France, 2006. Available from: 〈http://monographs.iarc.fr〉.
IRIS Progress Report, August 2011. 〈www.epa.gov/iris〉.
Jelcic, Z., 2004. Solvent molecular descriptors on poly (D, L-lactide-*co*-glycolide) particle size in emulsification–diffusion process. Colloids and Surfaces A: Physicochem. Eng. Aspects 242, 159–166.
Kar, S., Harding, A.P., Roy, K., Popelier, P.L.A., 2010. QSAR with Quantum Topological Molecular Similarity Indices: Toxicity of Aromatic Aldehydes to Tetrahymena pyriformis. SAR QSAR. Environ. Res. 21, 149–168.
Kar, S., Roy, K., 2010. Predictive toxicology using QSAR: A perspective. J. Indian Chem. Soc. 87, 1455–1515.
Kar, S., Roy, K., 2011. Development and validation of a robust QSAR model for prediction of carcinogenicity of drugs. Indian J. Biochem. Biophys. 48, 111–122.
Kar, S., Roy, K., 2012. First report on development of quantitative interspecies structure-carcinogenicity relationship models and exploring discriminatory features for rodent carcinogenicity of diverse organic chemicals using OECD guidelines. Chemosphere 87, 339–355.

Kier, L.B., Hall, L.H., Frazer, J.W., 1991. An index of electrotopological state for atoms in molecules. J. Math. Chem. 7, 229–241.
Klaassen, C.D., Watkins III, J.B., 1999. Casarett & Doull's Toxicology, Companion Handbook: The Basic Science of Poisons, 5th ed. McGraw-Hill, New York.
Massarelli, I., Imbriani, M., Coi, A., Saraceno, M., Carli, N., Bianucci, A.M., 2009. Development of QSAR models for predicting hepatocarcinogenic toxicity of chemicals. Eur. J. Med. Chem. 44, 3658–3664.
Matthews, B.W., 1975. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozime. Biochim. Biophys. Acta 405, 442–451.
MINITAB is a Statistical Software of Minitab Inc., USA, 〈http://www.minitab.com〉.
Mitra, I., Saha, A., Roy, K., 2010. Exploring quantitative structure-activity relationship (QSAR) studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. Mol. Simul. 36, 1067–1079.
Mitteroecker, P., Bookstein, F., 2011. Linear Discrimination, Ordination, and the isualization of Selection Gradients in Modern Morphometrics. Evol. Biol. 38, 100–114.
Müller, L., Kikuchi, Y., Probst, G., Schechtman, L., Shimada, H., Sofuni, T., 1999. ICH harmonized guidance on genotoxicity testing of pharmaceuticals; evolution, reasoning and impact. Mutat. Res. 436, 195–225.
Murcia-Soler, M., Pérez-Giménez, F., Garcıa-March, F.J., Salabert-Salvador, M.T., Dıaz-Villanuev, W., Medina-Casamayor, P., 2003. Discrimination and selection of new potential antibacterial compounds using simple topological descriptors. J. Mol. Graph. Model. 21, 375–390.
National Toxicology Program, 2005. U.S. Department of Health and Human Services. Public Health Service.
OECD Document, 2007. Guidance Document on the Validation of (Quantitative) 1226.
Structure-Activity Relationships (Q)SARs] Models, ENV/JM/MONO(2007)2.
Ojha, P.K., Mitra, I., Das, R.N., Roy, K., 2011. Further exploring r_m^2 metrics for validation of QSPR models. Chemom. Intell. Lab. Syst. 107, 194–205.
Perez-Garrido, A., Helguera, A.M., Borges, F., Cordeiro, M.N.D.S., Rivero, V., Escudero, A.G., 2011. Two New Parameters Based on Distances in a Receiver Operating Characteristic Chart for the Selection of Classification Models. J. Chem. Inf. Model. 51, 2746–2759.
Prado-Prado, F.J., Uriarte, E., Borges, F., Gonzaĺez-Dıáz, H., 2009. Multi-target spectral moments for QSAR and Complex Networks study of antibacterial drugs. Eur. J. Med. Chem. 44, 4516–4521.
Putz, M.V., Ionaşcu, C., Putz, A.M., Ostafe, V., 2011. Alert-QSAR. Implications for Electrophilic Theory of Chemical Carcinogenesis. Int. J. Mol. Sci. 12, 5098–5134.
Rogers, D., Hopfinger, A.J., 1994. Application of Genetic Function Approximation to Quantitative Structure Activity Relationships and Quantitative Structure Property Relationships. J. Chem. Inf. Comput. Sci. 34, 854–866.
Roy, K., Mitra, I., 2011. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Comb. Chem. High Throughput Screen 14, 450–474.
Roy, K., Mitra, I., Kar, S., Ojha, P., Das, R.N., Kabir, H., 2012. Comparative studies on some metrics for external validation of QSPR models. J. Chem. Inf. Model. 52, 396–408.
Schüürmann, G., Ebert, R.-U., Chen, J., Wang, B., Kühne, R., 2008. External validation and prediction employing the predictive squared correlation coefficient—test set activity mean vs. training set activity mean. J. Chem. Inf. Model. 48 (11), 2140–2145.
Shahlaei, M., Sabet, R., Ziari, M.B., Moeinifard, B., Fassihi, A., Karbakhsh, R., 2010. QSAR study of anthranilic acid sulfonamides as inhibitors of methionine aminopeptidase-2 using LS-SVM and GRNN based on principal components. Eur. J. Med. Chem. 45, 4499–4508.
SIMCA-P 10.0, 〈www.info@umetrics.com〉, UMETRICS, Umea, Sweden, 2002. 〈www.umetrics.com〉.
SPSS is a statistical software of SPSS Inc., USA. 〈http://www.spss.com〉.
STATISTICA is a Statistical Software of STATSOFT Inc., USA, 〈http://www.statsoft.com/〉.
Tanno, M., Sueyoshi, S., Miyata, N., Nakagawa, S., 1996. Nitric Oxide Generation from Aromatic N-Nitrosoureas at Ambient Temperature. Chem. Pharm. Bull. 44, 1849–1852.
Toropova, A.P., Toropov, A.A., Lombardo, A., Roncaglioni, A., Benfenati, E., Gini, G., 2010. A new bioconcentration factor model based on SMILES and indices of presence of atoms. Eur. J. Med. Chem. 5, 4399–4402.
US EPA, 2009. Integrated Risk Information System. US Environmental Protection Agency, National Center for Environmental Assessment, Washington, DC, Available from: 〈http://www.epa.gov/iris/〉 [accessed 1 January 2011].
USEPA IRIS database. 〈http://www.epa.gov/iris/〉.
Wang, N.C.Y., Venkatapathy, R., Bruce, R.M., Moudgal, C., 2011. Development of quantitative structure–activity relationship (QSAR) models to predict the carcinogenic potency of chemicals. II. Using oral slope factor as a measure of carcinogenic potency. Regul. Toxicol. Pharm. 59, 215–226.
Ward, J.M., 2010. Evolution of the uses of rats and mice for assessing carcinogenic risk from chemicals in humans. Asian Pac. J. Cancer Preven. 11, 18.
Williams, E.S., Panko, J., Paustenbach, D.J., 2009. The European Union's REACH regulation: a review of its history and requirements. Crit. Rev. Toxicol. 39, 553–675.
Wold, S., 1995. In: van de Waterbeemd, H. (Ed.), Chemometric Methods in Molecular Design. VCH, Weinheim, pp. 195–218.
Wold, S., Sjostrom, M., Eriksson, L., 1998. Partial least squares projections to latent structures (PLS) in chemistry. In: Schleyer, P. v.R. (Ed.), Encyclopedia of Comparative Chemistry, Vol. 3. Wiley, Chichester, GB.
Wold, S., Sjostrom, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130.